Presenting Uncertainty in Graphical Format

Edward Mulrow[1] and Nola du Toit[2]
[1]NORC at the University of Chicago, 4350 East-West Highway, F8, Bethesda MD 20814
[2]NORC at the University of Chicago, 55 E. Monroe, Suite 31, Chicago, IL 60603

Abstract
Presenting uncertainty in a chart or graph can be very challenging. A cursory scan on graphics in journals and reports indicates a wide variety of attempts to graphically present uncertainty, usually in the form of adding error bars or confidence intervals to point estimates. Using examples from our own past research on household structure among low-income families, we explore different graphical forms that may aid viewers in illustrating uncertainty. We also discuss common issues in creating and interpreting graphics that incorporate confidence intervals, margins of error, and hypothesis testing, and offer suggestions for best practices.

Key Words: Uncertainty, hypothesis testing, statistical graphics, violin plots, dirty plots.

1. The Problem with a Typical Graphic Showing Uncertainty

Typically, in statistics, the term "uncertainty" refers to the amount of error in the estimate of a population, such as a mean or median (Wainer 1996, 2000). That is, the greater the uncertainty, the less certain we are that an estimate presents the true value. Uncertainty can also refer to the amount of variation, noise, or abnormalities in data that make it deviate from what is expected. For example, the distribution of the data is skewed or it contains outliers. Finally, there is uncertainty in the conclusions we can draw based on the data. That is, we are able to declare groups as similar or different based on the overlap of the error or confidence interval (Wright et al 2019).

In addition to the abovementioned, there is also increasing uncertainty in the best methods of portraying error in graphics. Building on the work of others, (Wainer 1996, 2000; Wright et al. 2019; Hullman et al 2015), we provide guidance on how to present uncertainty in graphical format. This work is important as it addresses a common problem: many people do not know how to present or interpret uncertainty in graphics. With all the effort put forth to minimize uncertainty, and measure it accurately, more effort should be placed into the correct methods of publishing uncertainly so that it can be used appropriately (Gray, 2019).

In the sections below, we address these types of uncertainty and present the advantages and disadvantages of different graphical formats for each:

- Variation in the data
- Visualizing the uncertainty of estimates
- Visually determining potentially meaningful differences in the presence of uncertainty

1.1 Data

We use data from the third wave of the Making Connections Study (collected between 2008 and 2010). The longitudinal study collected data on low-income households in ten metropolitan cities.[1] We use the Median Household Income (in USD thousands) for all cases (n=1,619), as well as across different types of households (see Table 1). The households are categorized based on the coding of all adults in the home in relationship to a randomly selected focal child (du Toit, Bachtell, and Haggerty 2013).

Table 1: Sample Size and Percent of the Sample Population by Type of Households.

Type of Household	*Sample Size*	*Percent of Population*
Single Parent Only	518	26
Single Parent and at least one Grandparent	96	7
Single Parent and at least one Other Adult	156	10
Two Parents Only	586	39
Two Parents and at least one Grandparent	52	4
Two Parents and at least one Other Adult	64	4
Non Parent	147	10
Total	1,619	100

1.2 Illustrating the Problem

Figure 1 presents the median household income for different types of households. The graphic shows a common way of depicting uncertainty in graphics; horizontal bars that represent the estimate, with error bars plotting the confidence intervals.

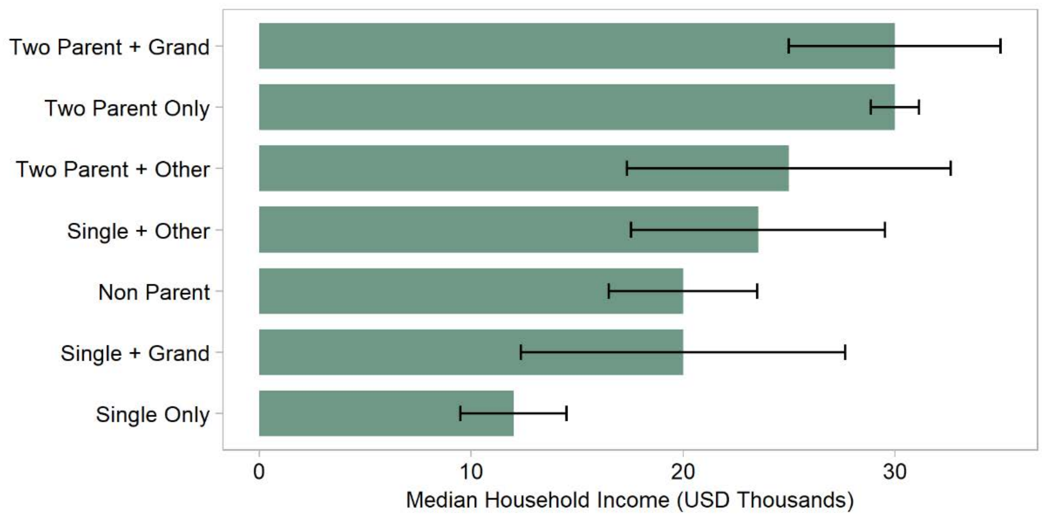


Figure 1: Example of typical graphic showing uncertainty. Visually it overemphasizes the point estimates to the detriment of the uncertainty. It should not be used.

[1] http://www.norc.org/Research/Projects/Pages/making-connections.aspx

Wainer (1996) would call it chart junk and wasteful:

> *"It is certainly profligate to use an entire bar when all of the information about the [estimate] is contained in the location of the top line; the rest is chart junk…"* (page 128)

The problem with this graphic is that is it trying to do too many things at once. First, bar charts are used to compare the length of bars across groups (Cleveland and McGill 1987). Considering that this graphic is presenting the median income – a single estimate for each group - the bars, however, place more emphasis on the estimates, instead of the uncertainty. Second, while the ends of the bars mark the estimates, the bars add clutter because they have to start at zero, thus deemphasizing the estimates and uncertainty. Third, the confidence intervals are presented. This may suggest that the graphic can be used to identify potentially meaningful differences by using the presence of overlapping bands.

Figure 1 has an overabundance of content and clutter. Consequently, we agree with Wainer that it is chart junk, and should not be used. The question is, then, how to best present uncertainty in graphics. Below we present a few examples of graphics for each of the three types of uncertainty listed above, along with the benefits and disadvantages of each type.

2. Exploring Variation in the Data

Considering that uncertainty can refer to variation in the data, there are many ways to visually explore the data. In doing this, we can examine the shape of the data, look for patterns and outliers, and even determine a model. Below are some techniques for exploring variation or uncertainty in the data.[2]

2.1 Boxplots

Box-plots plot summarizes the data and shows the distribution in a standardized way, with five points of interest: the minimum or lowest number or estimate, first quartile, median, third quartile, and maximum. With Box Plots, it is easy to see where middle 50% of the data are, judge symmetry versus skewness, and see outliers.

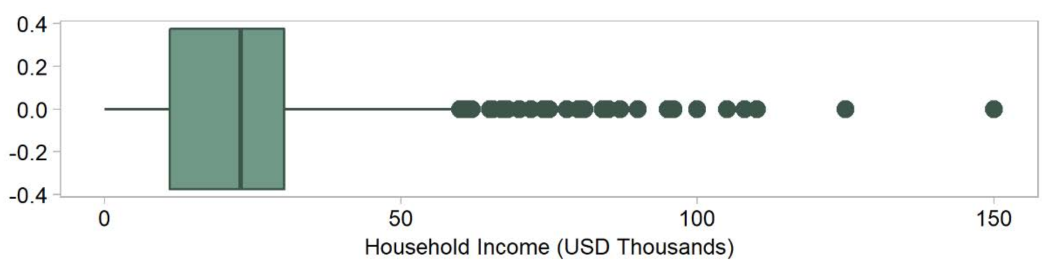


Figure 2: Boxplot of median income for all households.

Figure 2 shows the boxplot of all households' median income. The distribution of the data is clearly shown. Figure 3 shows the boxplots for each of the types of households. This format makes it easy to compare the median and distributions of the data across groups. However, it is less easy to see the shape of the data.

---

[2] These may not be appropriate plots for a general audience. Rather, they are informative for the analyst and will help to determine appropriate methods for analyzing the data.

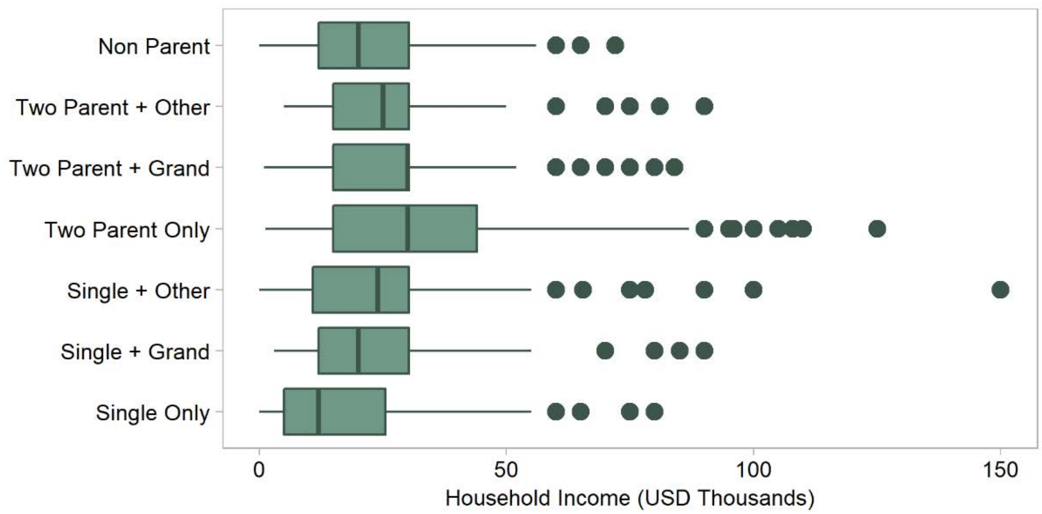


Figure 3: Boxplot of median income for each type of household.

2.2 Histograms

Histograms are used to judge the shape of the data. These displays are created by assigning a bin-width for each summary interval (buckets) for the density of the data. Figure 4 shows the median income data presented in a histogram. However, in order to find the best display, you need to experiment with the bin-width, along with the starting and stopping end points for the bins, to be sure you see the right shape of the data (Duong 2001).

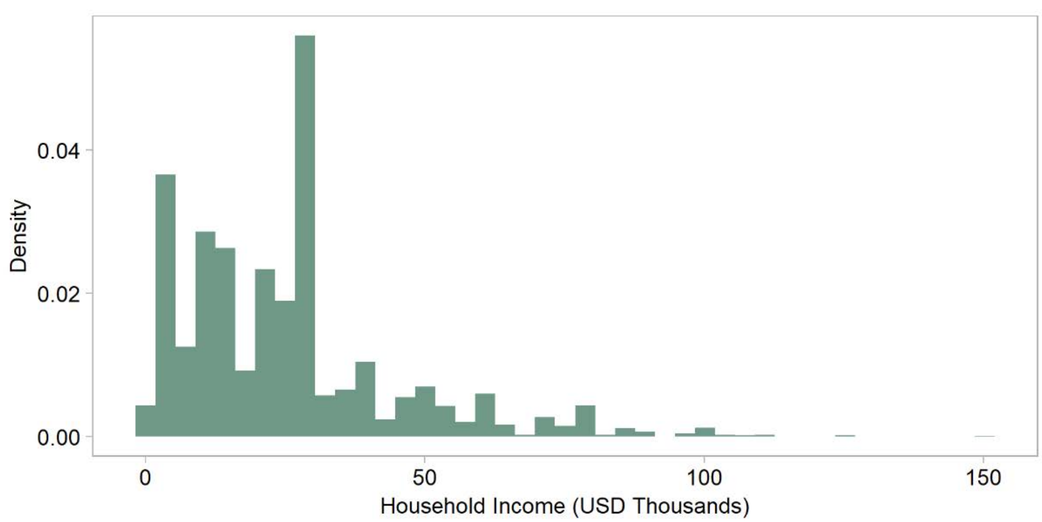


Figure 4: Histogram of median income for all households.

Figure 5 shows that the data looks different when assigning a different bin-width. It takes time to find the best bins, so this part of the process can be very explorative.

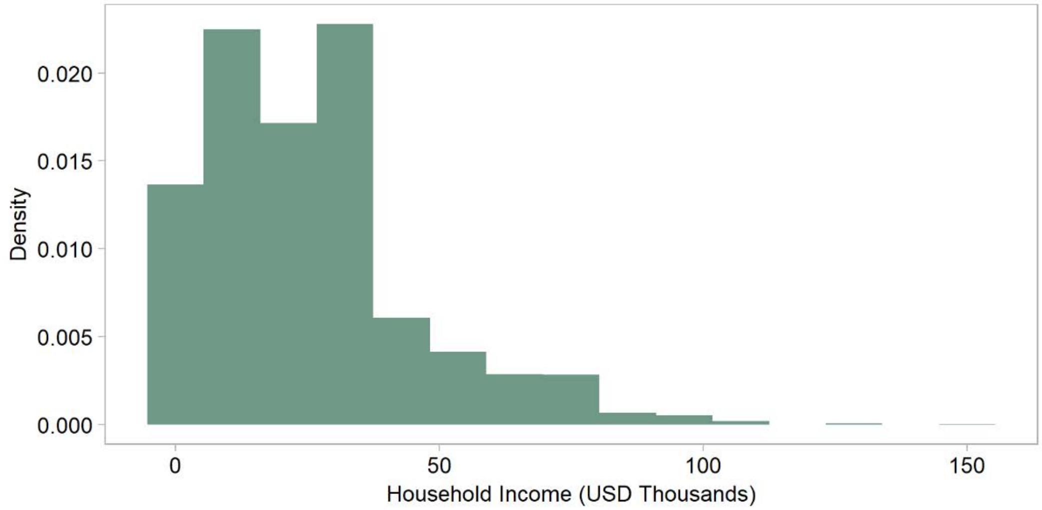


Figure 5: Histogram of median income with different bin-width.

Figure 6 presents histograms for each of the different family types. Unfortunately, histograms are not very good for comparisons across groups (Cleveland 1993). In order to best understand the shape of the data, each group may have to be assigned a different bin-width, and the choices for the starting and ending bins may vary.

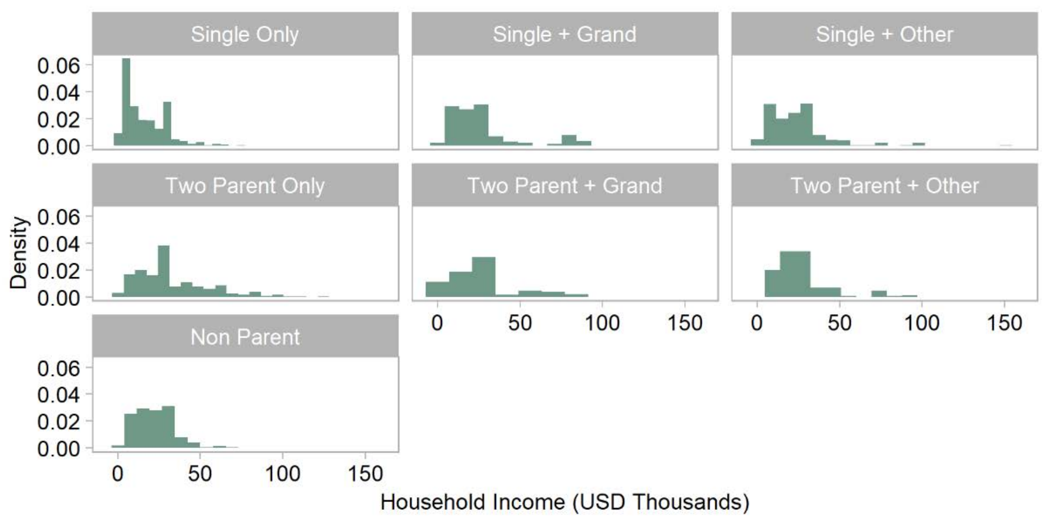


Figure 6: Histogram of median income for each type of household.

2.3 Density Plots

Density Plots, or Kernel Density Plots, are an alternative to the histogram, which solves the problem of choosing start and end points for the data display (Duong 2001). A kernel smoother is used to provide a continuous fit of a density function. This gives a reasonable visualization of the distribution shape. The methodology does depend on choosing a bandwidth, which is an analogue of bin width; however there has been extensive research that provides a means to choose an optimal bandwidth (Duong 2001). Figure 7 shows the density plot of median income for all households.

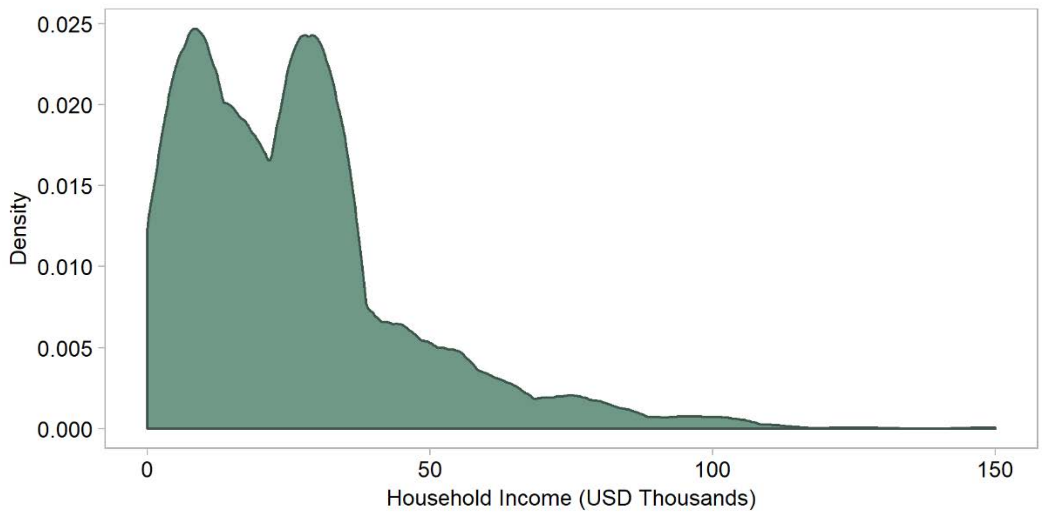


Figure 7: Density plot of median household income with bimodal distribution.

Figure 7 shows that the data is bimodal – that is, there are two modal values occurring most frequently, instead of one mode expected in a typical normal distribution. This illustrates the importance of exploring the data in various graphical formats. While the histogram in Figure 4 suggests that the distribution is not normal, it is greatly dependent on bin-width specifications; and the boxplots did not suggest bimodality at all.

2.4 Violin Plots

Violin plots combine elements from boxplots and density plots. That is, they plot the distribution of the data, as in the boxplot, but apply the kernelling of the density plots (Hintze and Nelson, 1998). Like the boxplots, they are good for comparison across groups. Figure 8 shows the median income for each type of household. Unfortunately, these plots are less known and rarely used to present data to the general public.

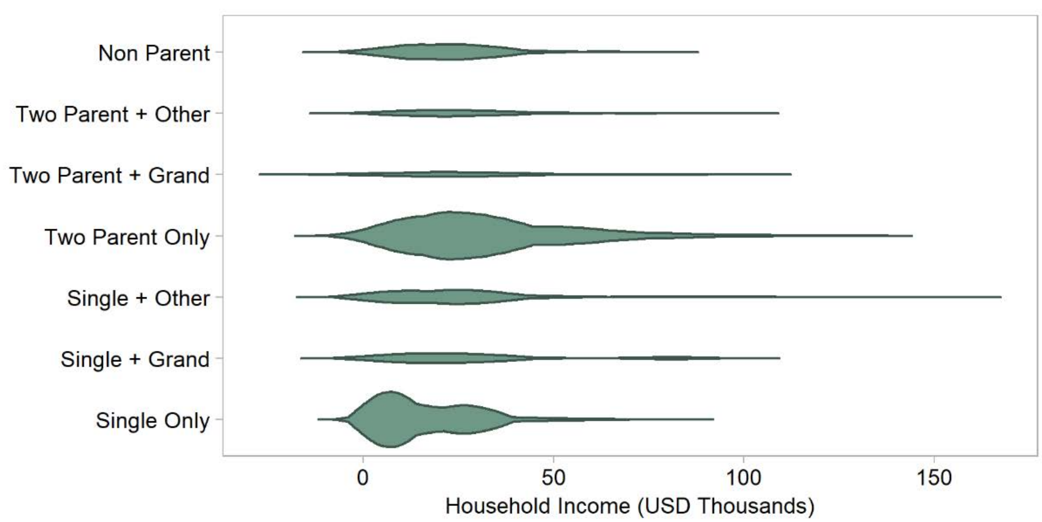


Figure 8: Violin plot of median income for each household type.

2.5 Other Plots to Try

We have presented a few of the more common ways of exploring variation or uncertainty in the data. But there are other plots that may be more suitable for your purposes. See below for references to alternative methods of plotting data.

- Raindrop, Rainforest, and Raincloud Plot
  - https://www.jstor.org/stable/30037295?seq=1#page_scan_tab_contents
  - https://mran.microsoft.com/snapshot/2017-12-15/web/packages/metaviz/vignettes/metaviz.html
  - https://peerj.com/preprints/27137v1.pdf

- Sectioned Density Plots
  - http://am.air.org/help/NAEPTextbook/htm/oSectDensPlot.htm

- Letter Value Boxplots
  - https://www.r-project.org/conferences/useR-2006/Slides/HofmannEtAl.pdf

3. Visualizing the Uncertainty of Estimates

Estimates are derived from an observed sample. Other researchers using different samples will likely derive different estimates. The uncertainty we want to visualize is sample-to-sample variation for estimates of a characteristic (parameter). In a table, estimates and their standard errors or margins of error are listed. It is not as easy to present uncertainty in graphical format. In general, we try to show the uncertainty of an estimate by visualizing the estimate and a confidence interval (CI).

3.1 Dot Plots

Dot plots are a common way of presenting estimates and their uncertainty. They present the point estimate with a dot (other symbols can also be used), and typically show the confidence interval or error bar surrounding the estimate using a line in order to illustrate the level of uncertainty in the estimate. Figure 9 shows the median household income for the different types of households, using dots for the point estimate and lines for the 95% CI.

We determined the CI using this formula:

$$\text{Upper bound} = \text{Household income} + 1.96 * se$$
$$\text{Lower bound} = \text{Household income} - 1.96 * se$$

While many believe Figure 9 is superior to Figure 1, the display may not provide the feeling of uncertainty one might like if showing uncertainty is all that is intended. Points and lines provide the viewer a more precise way to view the data. Cleveland and McGill (1987) refer to this as good table lookup. The 95% CI, with such stringent upper and lower bounds (+ or – 1.96 of the standard error), suggest that uncertainty has a starting and stopping point. This can lead to misinterpretations of the level of uncertainty (Cairo, 2019). Viewers may also misinterpret overlapping CI. We will discuss that in Section 4. For now, we consider a visualization that uses plotting methods that provide the viewer less certainty. We suggest the use of dirty plots, also called sloppy or fuzzy plots (Wainer 1996; Jackson, 2008; Cairo, 2019).

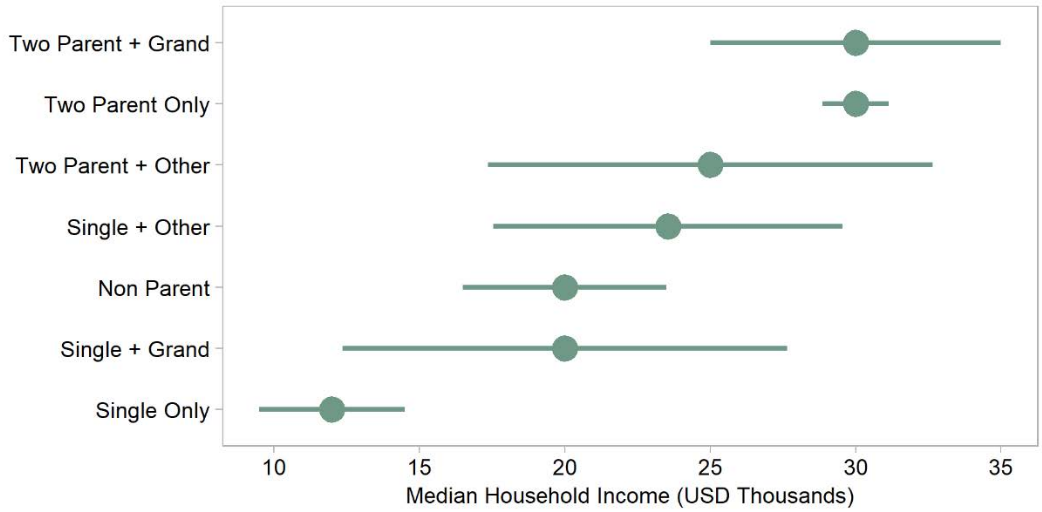


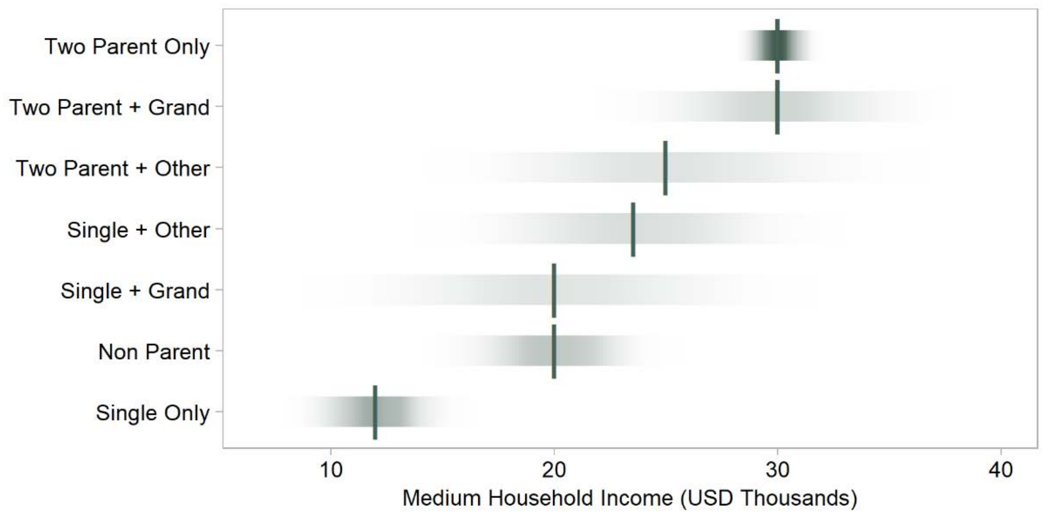
Figure 9: Dot plot of estimates with 95% CI presented with lines.

3.2 Dirty Plots

Dirty Plots are a less common way of presenting uncertainty in estimates. These present the full density of the estimates distribution in a shaded band, instead of a solid line, thus making it more difficult to show (1) the point estimate, and (2) where the uncertainty starts or ends. As compared to the solid CI lines on Figure 9 that suggest specific starting and stopping points, the graded shading creates the feeling of variability or uncertainty. While it is difficult to find potentially meaningful differences, it makes it more apparent that we are not clear on the true estimate for the sample or the level of uncertainty (see Figure 10).

Figure 10: Dirty plot of median income for each type of household. There is no dot for the estimate and bands with graded shading have been added for the CI.

Figure 11 includes the addition of the point estimate, but still illustrates the importance of showing the uncertainty. Ultimately, dots and lines are more exact, but shading suggests uncertainty – and if that is the point, then dirty plots are preferable.



.
Figure 11: Dirty plot of median income for each type of household, with bars added for the estimates.

4. Determining Potentially Meaningful Differences in the Presence of Uncertainty

The identification of a potentially meaningful difference in the presence of uncertainty can be aided by well-developed visualizations. However, caution is needed when using CI for visually testing the hypothesis that there is a difference between two group characteristics (e.g. mean, median, etc.). For example, we can determine that the median income for Single Parent Only and Two Parent Only in Figure 9 are potentially different in a meaningful way because their CI do not overlap. However, for the groups in which the bands do overlap, we need to do more investigating.

4.1 Illustrating the Problem

Figure 12 is a dot plot that shows the median income and 95% CI for Missouri and North Carolina.[3] The CI overlap, but we cannot not yet draw any conclusions. We need to consider the estimate of the difference and the corresponding CI.

4.2 Testing the Difference

In order to test if one state has a higher (or lower) median income than another, we determine whether the difference between the two medians is different from zero. The standard error of the difference is the square root of the sum of the squared standard errors of each estimate:

$$S_{diff} = \sqrt{S_1^2 + S_2^2}$$

[3] U.S. Census Bureau, 2017 American Community Survey.

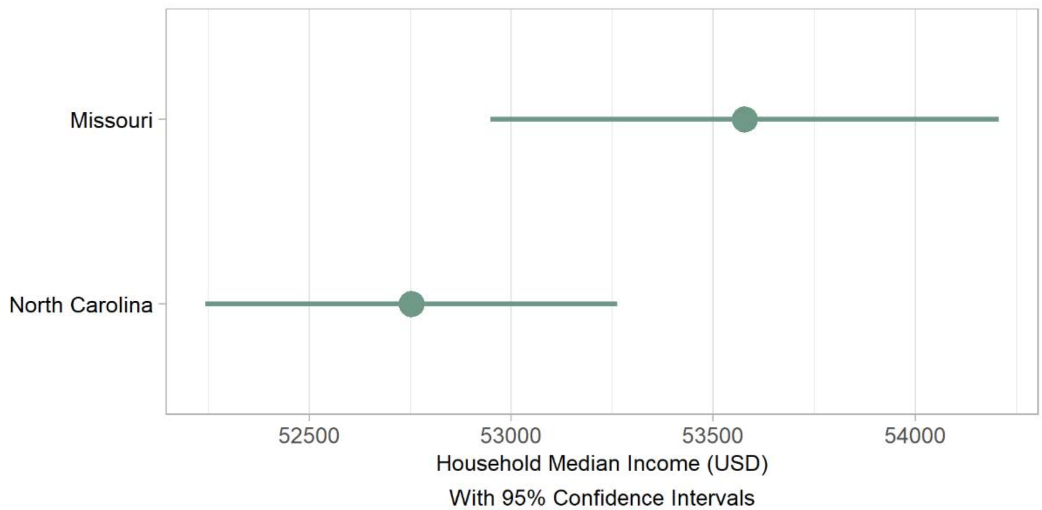


Figure 12: Dot plot of median household income for Missouri and North Carolina with 95% CI.

Figure 13 shows the estimate of the difference and its 95% CI. The confidence interval for the difference does not include 0. That is, the CI do not cross the vertical zero line. Some researchers would say that is evidence that we should reject the null hypothesis that the household median income estimates are the same.

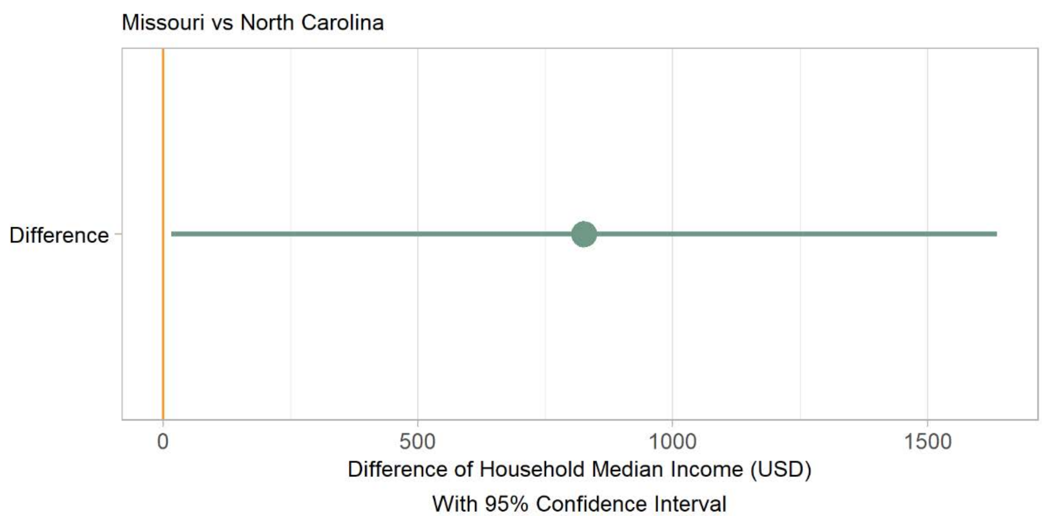


Figure 13: Estimate for the difference with 95% CI. The band does not cross vertical zero line and we reject the null hypothesis that there is no difference.

4.3 Alternative to Graphing the Difference

Some researchers may still want to see each estimate separately. In that case, we can adjust the dot plot display by using the margin of error for the difference with half of it added and subtracted to each estimate.

Figure 14 is a dot plot with the estimate for each state. It also shows the margin of error at the 95% confidence level for the difference (1.96 * S_{diff}); half of the difference margin of error has been added to and subtracted from each dot.

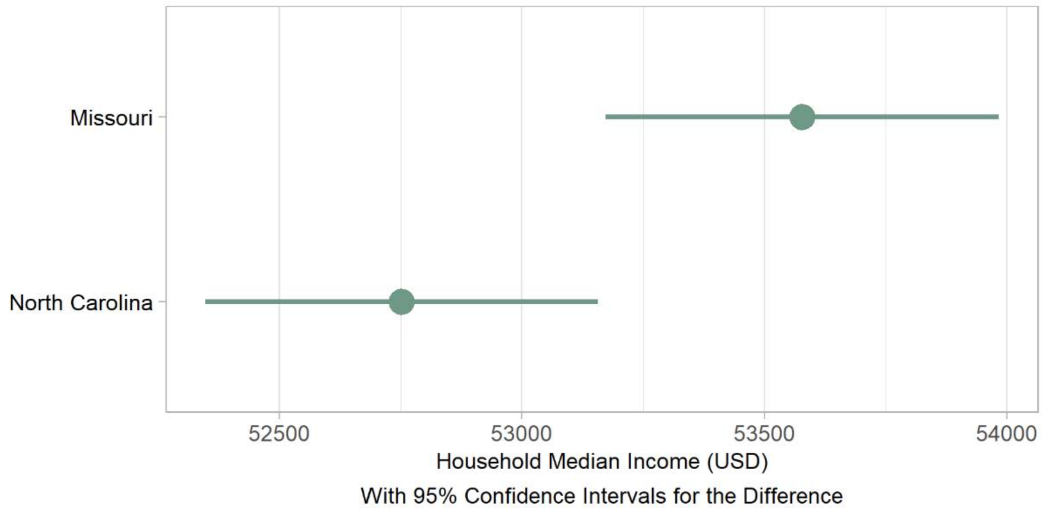


Figure 14: Plotting the Estimates with half the difference margin of error subtracted/added from/to each estimate.

Figure 14 shows that there is no overlap in the 95% CI. So there is reason to believe there is a potentially meaningful difference between the median incomes of Missouri and North Carolina. For more illustration, Figure 15 shows the estimates with the two sets of CI - the adjusted CI are shown with the dotted lines.

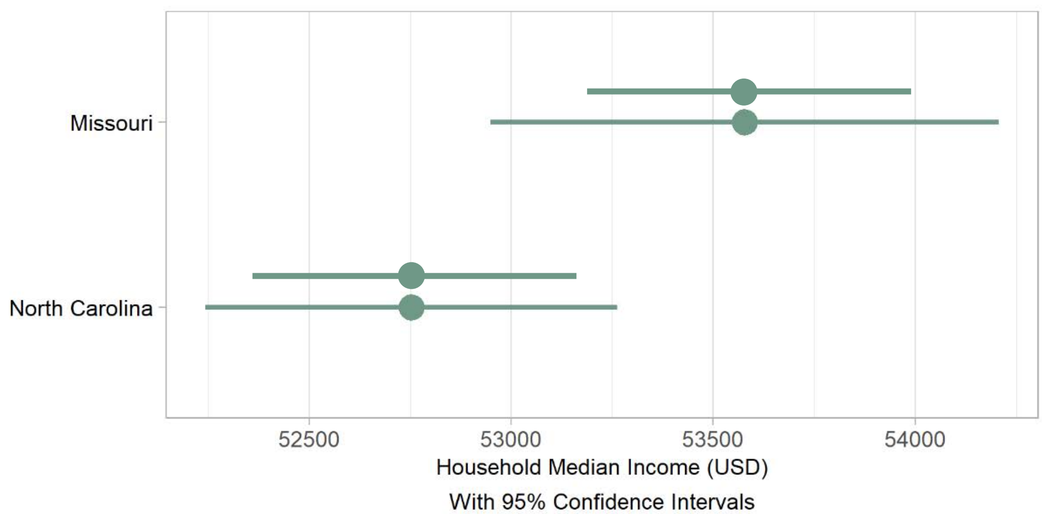


Figure 15: The original estimates and CI are plotted, along with the adjusted CI (dotted lines).

A word of caution here: when there are more than two estimates in the plot, the situation is more complex. There can be numerous pairs to test. If there are three point estimates, there are three differences to consider. How do you choose one margin of error to plot around each estimate? The more estimates plotted, the worse it gets. For four estimates, there are (4*3)/2 = 6 pairwise comparisons. For the 50 states plus DC there are (51*50)/2 = 1,275 pairwise comparisons!

In addition to the issue of the number of pairwise comparisons a viewer could make, there is also the problem of multiple testing. Due to uncertainty, there is a chance in any single hypothesis test that you observed an unusual sample that leads you to the wrong conclusion. When there are many tests taking place at the same time, as there would be if many estimates are plotted with CI, the changes of drawing the wrong conclusion increases.

Many authors such as Wainer (1996), Almond et al 2000, and, most recently, Wright et al (2019) have provided advice on how to adjust margins of error to so that more than two estimates can be plotted with CI, so that the viewer can perform visual hypothesis testing by looking for non-overlapping CI. If the purpose is to create a plot so that a viewer can choose any two estimates to test, a conservative margin of error can be found for each estimate to judge pairwise differences (Wainer 1996; Wright et al. 2019). This includes using a Bonferroni corrections to adjust for multiple comparisons.

Figure 16 was created using R RankingProject package. It provides a conservative visual approach to comparing pairwise median income differences for all 50 state, DC and Puerto Rico. If any two CI do not overlap, it is safe to assume, with at least 95% confidence, that there is a potentially meaningful difference between the median incomes of the two regions.

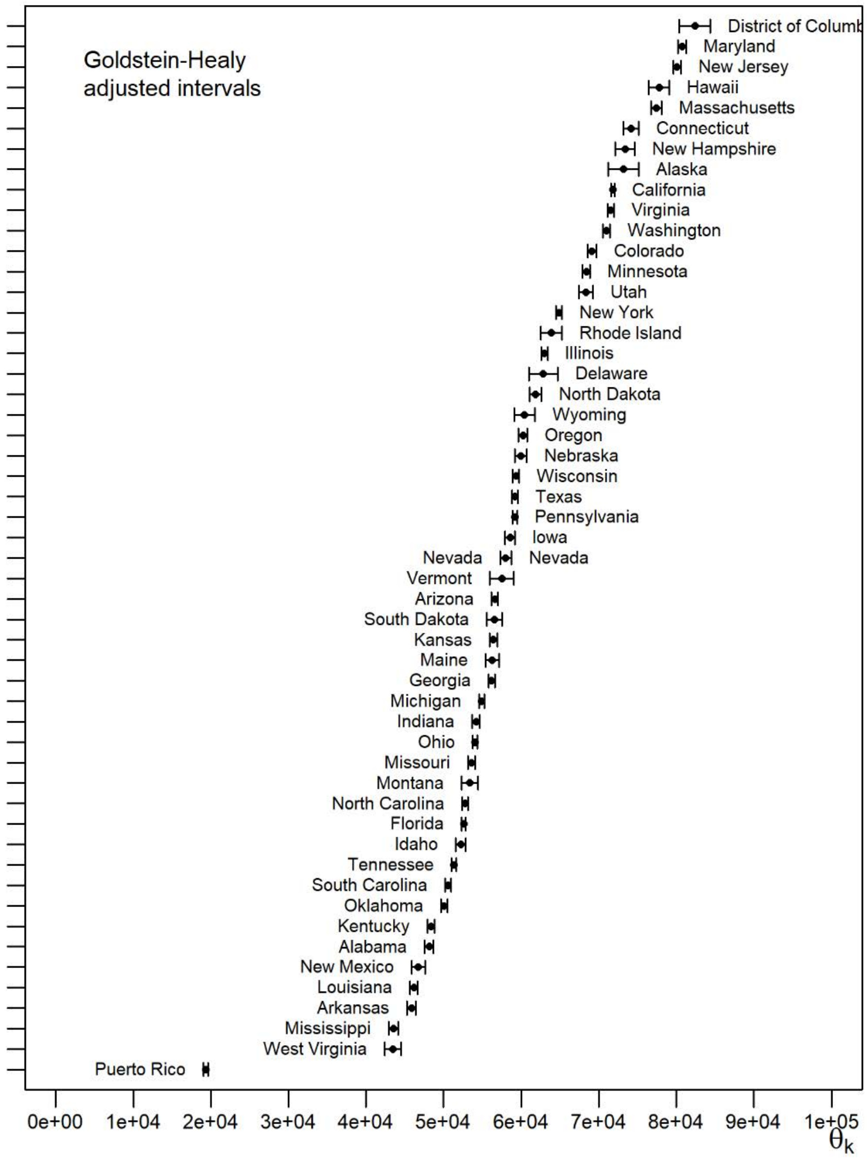


Figure 16: Median Income for all States, District of Columbia, and Puerto Rico, with adjusted CI. Data come from the US Census Bureau 2017 American Community Survey.

Almond et al. (2000) point out that in many situations a researchers may want to compare one reference estimate to all others. In this case you can create a more focused graphic. Using the state median income data, Colorado is set as the reference state (as JSM 2019 was in Colorado). Figure 17 uses the R RankingProject package to create a graphic for such a comparison. The graphic is designed so that if another state's CI do not overlap with the shaded band around Colorado's median income estimate, then one should note that there is potentially meaningful difference between the median income of that state and Colorado. See Wright, et al (2019) for the details.

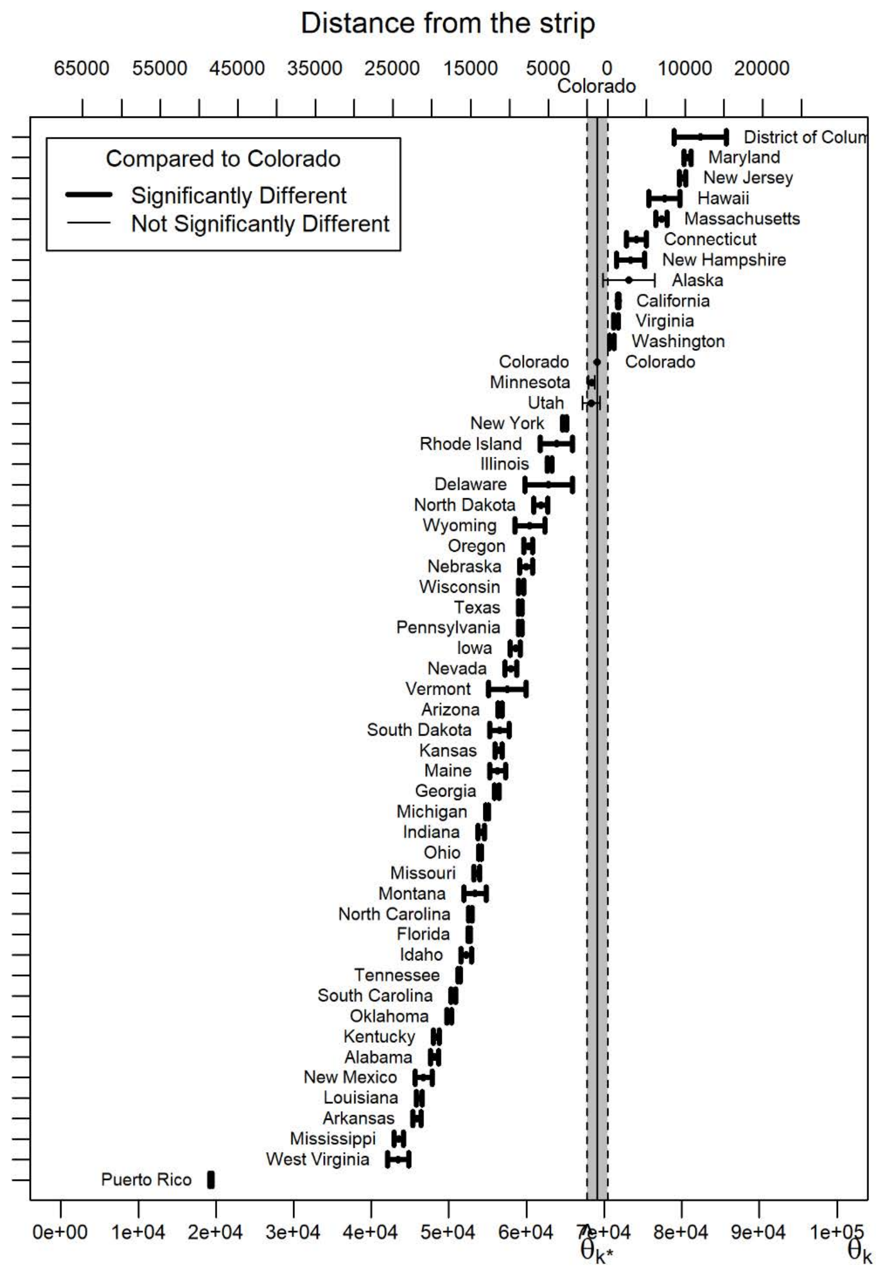


Figure 17: Pairwise Differences for Median Income for all States, District of Columbia, and Puerto Rico, Compared to Colorado. Potentially meaningful differences exist between the median income of Colorado and any other state whose median income CI does not overlap with the shaded CI region for Colorado.

5. Final Notes

Uncertainty can be difficult to incorporate into visualization. Not all graphics need to have uncertainty incorporated – it depends on the story. If a lot of time and effort has gone into determining the uncertainty of estimates, time and effort should go into visualizing the uncertainty. Uncertainty should be incorporated, in some way, if the graphics are exploratory in nature. Some display methods convey certainty, and should be avoided if the main point is to convey uncertainty. However, when presenting to others, consider, does the audience really need to see uncertainty?

References

Almond, R. G., Lewis, C., Tukey, J. W., and Yan, D. 2000. "Displays for Comparing a Given State to Many Others." *The American Statistician* 54:89-93.

Cairo, A. 2019. *How Charts Lie*. ISBN-13: 978-1324001560.

Cleveland, W. 1993. *Visualizing Data*. Hobart Press, Summit NJ.

Cleveland, W. and McGill, R. 1987. “Graphical Perception: The Visual Decoding of Quantitative Information on Graphical Displays of Data.” *Journal of the Royal Statistical Society, Series A* 150:192-229.

Duong, T. 2001. *An introduction to kernel density estimation*, University of Western Australia, http://www.mvstat.net/tduong/research/seminars/seminar-2001-05/index.html

du Toit, N., Bachtell, K., and Haggerty, C. 2013. Coming and Going: Effects of Change in Household Composition on the Economic Wellbeing of Families with Children. In C. C. Yeakey, V. Sanders-Thompson & A. Wells (Eds.), Urban Ills: Post Recession Complexities of Urban Living in Global Contexts, New York, NY: Lexington Books.

Gray, D. 2019. Personal communication.

Hintze, J. L. and Nelson, R. D. 1998. “Violin Plots: A Box Plot-Density Trace Synergism,” *The American Statistician* 52: 181-184.

Hullman, J., Resnick, P., and Adar, E. 2015. “Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences About Reliability of Variable Ordering.” *PLOS ONE* 10:11. http://idl.cs.washington.edu/papers/hops.

Jackson, C. 2008. “Displaying Uncertainty with Shading,” *The American Statistician* 62: 340-347.

Wainer, H. 1996. "Depicting Error," *The American Statistician* 50:101- 111.

Wainer, H. 2009. Picturing the Uncertain World: How to Understand, Communicate, and Control Uncertainty through Graphical Display. Princeton, NJ: Princeton University Press.

Wright, T., Klein, M., Wieczorek, J. 2019. "A Primer on Visualizations for Comparing Populations, Including the Issue of Overlapping Confidence Intervals." *The American Statistician* 73:2, 165-178. DOI: 10.1080/00031305.2017.1392359.